Fold and function predictions for *Mycoplasma genitalium* proteins

Leszek Rychlewski, Baohong Zhang and Adam Godzik

Background: Uncharacterized proteins from newly sequenced genomes provide perfect targets for fold and function prediction.

Results: For 38% of the entire genome of *Mycoplasma genitalium*, sequence similarity to a protein with a known structure can be recognized using a new sequence alignment algorithm. When comparing genomes of *M. genitalium* and *Escherichia coli*, > 80% of *M. genitalium* proteins have a significant sequence similarity to a protein in *E. coli* and there are > 40 examples that have not been recognized before. For all cases of proteins with significant profile similarities, there are strong analogies in their functions, if the functions of both proteins are known. The results presented here and other recent results strongly support the argument that such proteins are actually homologous. Assuming this homology allows one to make tentative functional assignments for > 50 previously uncharacterized proteins, including such intriguing cases as the putative β-lactam antibiotic resistance protein in *M. genitalium*.

Conclusions: Using a new profile-to-profile alignment algorithm, the three-dimensional fold can be predicted for almost 40% of proteins from a genome of the small bacterium *M. genitalium,* and tentative function can be assigned to almost 80% of the entire genome. Some predictions lead to new insights about known functions or point to hitherto unexpected features of *M. genitalium*.

Address: Department of Molecular Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA 92037, USA.

Correspondence: Adam Godzik
E-mail: adam@scripps.edu

Key words: fold prediction, function prediction, genome analysis

Received: **12 February 1998**
Revisions requested: **12 March 1998**
Revisions received: **26 March 1998**
Accepted: **02 April 1998**

Published: **26 May 1998**
http://biomednet.com/elecref/1359027800300229

Folding & Design 26 May 1998, **3**:229–238

Introduction

Recent years have brought a rapid increase of pace in determining new protein sequences. At present, over 300,000 protein sequences are deposited in public databases and several entire genomes of mostly prokaryotic organisms are known. Most of the new sequences are known only at the cDNA level. A computer analysis of their sequences is a primary source of information about their function and structure.

In particular, if homology to an already characterized protein family can be established, it is possible to make various inferences about the new protein structure, activity and function. Studying the sequence similarity between two proteins can be used to argue for or against their homology. For any measure of similarity, it is possible to determine the distribution of scores between unrelated proteins. Then, it is possible to calculate the E value, the number of proteins with a given score that could be expected by chance. A very small E value is usually taken as an argument that the score could not have happened by chance and the two proteins being compared are therefore homologous.

The functional and structural prediction by homology to already characterized proteins is extremely successful, being fast, inexpensive and reliable. There are several publicly available programs, such as BLAST [1] or FASTA [2], as well as many commercial or public software packages [3,4] or World Wide Web services [5,6] geared toward recognition of protein homology by the analysis of sequence similarities. Unfortunately, all such programs fail to recognize unrelated proteins that have a fold similar to that of an already known protein. They also fail for distantly related proteins when the sequence similarity drops to the level of random similarity between unrelated proteins.

Two different sets of tools were developed to address these two seemingly different problems. Superseding and/or enhancing the sequence–sequence similarity by sequence–structure compatibility allowed searching for unrelated proteins with similar structures [7–11]. Using additional information from multiple alignments of already identified homologous proteins extended the application of sequence alignment tools to recognize distantly related proteins [12,13].

These two approaches ask two seemingly different questions and strive to achieve apparently different goals. The first approach, usually referred to as threading, strives to match a sequence to a structure, targeting proteins with a similar three-dimensional structure with or without any homology between them [7–11]. The second approach uses sequences of closely related proteins to estimate the patterns of mutations along the sequence and to create the position-specific mutation matrix [12,13]. The objective of this approach is the same as in the standard sequence alignment methods — to identify homologies between proteins, or in this case protein families. Thus,

in principle, the threading approach has a much wider application than the profile, or any other sequence-only type approach. A significant limitation of threading is that it can be used only for proteins with known structures. On the other hand, sequence-based methods seeking to recognize homology between proteins can use proteins for which the structure is not known, and at the same time, they can achieve much more than just structure prediction. If a protein can be placed into the already characterized family of homologous proteins, there might be features, other than the structure, that are shared by all or some members of the family. Features such as function, activity, mechanism of action, cofactors necessary for their activity and general patterns of interactions with other molecules are often shared by homologous proteins, depending on the evolutionary distance between them.

Recently, we have demonstrated that the profile alignment methods can closely match the threading algorithms (see the Materials and methods section for an example) in terms of recognizing fold similarity [14]. This strongly supports the growing realization that in most known examples of apparently unrelated proteins with similar structures, the proteins in question are actually homologous. This is in direct contradiction to the ideas expressed in the early threading papers, which explicitly claimed going beyond the 'homology paradigm' [9]. Detailed analysis of many specific pairs and groups of structurally similar proteins suggested that they represent very distant homologs [15–18]. At the same time, more sophisticated methods of sequence comparison, such as Hidden Markov Models [19] and PSI-BLAST ([20]; discussed later in this paper) extended the reach of sequence similarity recognition into the region, which before was thought to represent purely random sequence similarities. All these results open a new interpretation of fold predictions.

According to this interpretation, methods such as threading or other analogy-based structure prediction methods do not actually predict a fold, but instead, they recognize the homologous family to which the new protein belongs. The fold similarity between a new protein and a known structure from the family is simply a consequence of their homology. As for other features, their predictions are limited only by the evolutionary distance between the protein being predicted and the already characterized protein family and by our understanding of how such features change in homologous families. Unfortunately, we do not have a deep understanding of how functions evolve in homologous families. Most of our understanding is based on the analysis of relatively closely related proteins and/or on families of orthologous proteins, in which the function remains the same for all proteins. Thus, all function assignments from homology are tentative and subject to verification and, possibly, significant change.

Recently, a hybrid sequence threading algorithm was applied to the fold assignment for proteins of *Mycoplasma genitalium* [21]. 103 proteins could be assigned to a three-dimensional fold; 28 more than a standard sequence-based approach [21]. Here, we re-examine the proteins using two sequence-only tools, geared by design to recognize homologies between protein families. PSI-BLAST [20], the newly improved rapid database search algorithm BLAST, is a 'state-of-the-art' sequence similarity tool. BASIC (bilateral amplified sequence information comparison) is a profile–profile alignment method from a fold recognition suite developed in our group [14,22].

M. genitalium is a small, pathogenic Gram-positive bacterium associated with pulmonary and urogenital infections in humans. Its close cousin, *Mycoplasma pneumoniae* causes primary atypical pneumonia. *M. genitalium* is a very simple organism lacking a cell wall and is the smallest known cellular organism capable of independent replication. Its entire genome, composed of 468 predicted open reading frames, was sequenced by 'shotgun' sequencing [23] and made available on the World Wide Web site for The Institute for Genome Research (www.tigr.org) with annotations identifying the function of ~67% of all proteins in this genome.

Both PSI-BLAST and BASIC are used here in two different tasks. In the first part of the Results section, structural predictions for proteins from the *M. genitalium* genome are made by comparing their sequences against sequences and sequence profiles of proteins with known structures. In this application, both programs are used merely as fold prediction tools. The analysis of structure prediction results is used to argue for a broader interpretation of the prediction results as tentative assignments of *M. genitalium* proteins to homologous superfamilies. In this spirit, the same proteins are compared against proteins from the *E. coli* genome to study how many functional assignments can be transferred between genomes.

Results

Structural predictions

The set of 468 protein sequences from *M. genitalium* genome was downloaded from the World Wide Web site for The Institute for Genome Research. Each of these sequences was compared to a large protein sequence database using the PSI-BLAST [20] algorithm. In the next step, the same sequences were compared to a smaller database containing sequence profiles of a set of proteins representing all currently known protein folds using a profile–profile alignment program BASIC, a part of the suite of fold-prediction algorithms [14,22]. Technical details about the algorithms, databases and protocols for fold assignments are discussed in the Materials and methods section.

For the 468 protein sequences, the PSI-BLAST algorithm detected 126 significant (i.e. the number of proteins with

the same score expected by chance, the E value, < 0.1) similarities to proteins with known structures. This constitutes 27% of the entire genome, the ratio being much higher than the 10% [24] or 16% [21] reported previously. The increase partly results from the increase in the number of known structures, but mostly it can be attributed to the much greater sensitivity of a new generation of BLAST algorithms. For comparison, for the same genome, Fischer and Eisenberg [21] reported 75 significant similarities using the older version of the BLAST algorithm and a smaller version of the database of known structures.

The BASIC program from our fold-recognition suite detected 176 significant (E value < 0.05) similarities to proteins with known structures (38%), an almost 40% increase over the PSI-BLAST recognition rate. This is a superset of BLAST predictions because all the high-significance BLAST predictions are independently recognized by the BASIC algorithm. Thus, there are 50 new structural assignments, which are listed in Table 1. The significance threshold of the E value of 0.05 (see the Materials and methods section and the discussion in the next paragraph) is rather conservative, so there is a good chance that many predictions with a lower significance level are actually accurate. The complete analysis of the entire list of all 468 predictions is available from the authors' World Wide Web site (http://cape6.scripps.edu).

As discussed in the Materials and methods section, as a result of different databases and score distributions, significance assignment is difficult to compare between different methods, despite being expressed in a way analogous to that of the E value. Much more relevant are significance values of predictions that are known to be wrong. Such values could be used to 'calibrate' the method. For instance, the MG412 protein is predicted by BASIC to be similar to tyrosine phenol lyase (PDB code 1tpl) with the E value of 1×10^{-10} and by PSI-BLAST to be similar to glucosamine phosphate synthase (PDB code 1gdo) with the (above threshold) E value of 0.2 (Table 1). Both predictions could not be correct at the same time, because the folds of 1tpl and 1gdo are different. The BASIC prediction results in a reasonable model with a well-conserved cofactor (pyridoxal-5′-phosphate, PLP) binding site and a predicted binding site compatible with its function (methyltransferase). At the same time, the BLAST prediction results in an alignment with unphysically long gaps and the resulting model has no active site. This strongly argues that the BLAST prediction is wrong and, thus, the significance threshold for the E value of 0.1, recommended by the authors of PSI-BLAST [20], could not be lowered significantly. The appropriate significance threshold for the BASIC algorithm remains to be tested. The lowest E value for a known false prediction found during the testing and development of the algorithm was equal to 1.6. To account for possible differences between distributions on the testing set and the full prediction set, we used a very conservative E value of 0.05. More extensive testing of the BASIC algorithm may allow this value to be lowered and thus increase the number of fold predictions.

Functional analysis of structural predictions

Verification of fold predictions such as those presented in Table 1 is difficult because none of the structures is known. Several *M. genitalium* proteins in Table 1, such as arginyl-tRNA synthetase (MG382), phenylalanyl tRNA synthetase (MG194), cysteinyl-tRNA synthetase (MG253) and trptophanyl-tRNA synthetase (MG127), uridine kinase (MG382) or uracil phosphoribosyltransferase (MG030), have obvious functional similarities to the proteins that were identified by the BASIC algorithm. Homologies between tRNA synthetases were postulated previously based on conserved short patterns around the active site [25]. Taken together, these observations can be used as a strong argument that these pairs are indeed homologous. For the 29 hypothetical *M. genitalium* proteins in Table 1, no such indirect verification is possible because their function is not known. In two other examples (MG340 and MG218), only one domain from a large multidomain protein could be identified, so again even an indirect verification is not possible. In the remaining predictions, the functions of *M. genitalium* proteins are known, making it possible to discuss the predicted homology from this point of view. Of course, such arguments could not be used to verify the structural predictions because there are many examples of functional divergence between homologous proteins and functional convergence of non-homologous proteins. For many examples from Table 1, however, the functional similarity does support, or at least does not contradict, the possible homology between the pairs recognized by the BASIC algorithm. Eight examples are now given.

First, ribosomal protein S12 (SWISS-PROT code RS12_MYCGE, MG087) was predicted to have the OB-fold, already seen in several ribosomal and DNA/RNA binding proteins. It is interesting to note that a protein that was recognized as a homolog, the translational initiation factor IF1 from *E. coli*, is involved in ribosome binding. The S12 ribosomal protein is known to bind RNA directly and to interact strongly with other ribosomal proteins [26]. This may shed some light on the presently unknown mechanism of IF1 [27]. Only one 80 amino acid domain from the 139-residue S12 protein can be predicted. The remaining part is predicted to be predominantly helical.

Second, the nitrogen fixation (NIFS) protein homolog (SWISS-PROT code NISH_MYCGE, MG335) is predicted to be homologous to the pyridoxal phosphate dependent transferase superfamily, which includes aspartate and pyruvate aminotransferases, tyrosine phenol lyase and dialkylglycine and ornithine decarboxylase [28]. Distant homology between the NIFS family and pyruvate aminotransferase

Table 1

Novel fold assignments obtained with the BASIC algorithm.

				BASIC				PSI BLAST		
MG	Length	Target name		E value	Length	PDB	Template name	PDB	E value	A
364	679	Y364_MYCGE	Hypothetical protein MG3	0.05	224	1bgw_	Topoisomerase	–	> 10	
340	297	RPOC_MYCGE	DNA-directed RNA polymer	0.05	1292	1enp_	Enoyl acyl carrier protein	1bcp	6.8	N
380	214	GIDB_MYCGE	Glucose inhibited division	0.03	192	1vid_	Catechol o-methyltransferase	–	> 10	
183	305	PEPF_MYCGE	Oligoendopeptidase F	0.03	607	1sig_	RNA polymerase primary sigma	–	> 10	
147	514	Y147_MYCGE	Hypothetical protein	0.03	375	1occA	Cytochrome *c* oxidase	–	> 10	
397	679	Y397_MYCGE	Hypothetical protein	0.03	566	1bgw_	Topoisomerase	–	> 10	
206	271	UVRS_MYCGE	Excinuclease ABC chain C	0.03	432	1exnA	5′-exonuclease (5′-nuclease)	1taq	2.7	N
075	679	Y075_MYCGE	Hypothetical protein	0.03	1024	1bgw_	Topoisomerase	–	> 10	
414	679	Y414_MYCGE	Hypothetical protein	0.03	1036	1bgw_	Topoisomerase	–	> 10	
382	196	URK_MYCGE	Uridine Kinase	0.03	213	1ukz_	Uridylate kinase complexed	1ukz	0.3	Y
123	305	Y123_MYCGE	Hypothetical protein	0.03	471	1sig_	RNA polymerase primary sigma	1sig	2.1	Y
293	214	Y293_MYCGE	Hypothetical protein	0.02	244	1vid_	Catechol o-methyltransferase	–	> 10	
210	514	LSPA_MYCGE	Putative Lipoprotein	0.02	181	1occA	Cytochrome *c* oxidase	–	> 10	
213	679	Y213_MYCGE	Hypothetical protein	0.02	471	1bgw_	Topoisomerase fragment	–	> 10	
096	305	Y096_MYCGE	Hypothetical protein	0.02	527	1sig_	RNA polymerase primary sigma	–	> 10	
374	468	SYR_MYCGE	Arginyl-tRNA synthetase	0.02	537	1gln_	Glutamyl-tRNA synthetase	–	> 10	
112	329	Y112_MYCGE	Hypothetical protein	0.02	209	1ak5_	Monophosphate dehydrogenase	1noy	3.9	N
264	186	Y264_MYCGE	Hypothetical protein	0.01	198	1gky_	Guanylate kinase	–	> 10	
148	679	Y148_MYCGE	Hypothetical protein	0.01	409	1bgw_	Topoisomerase	–	> 10	
011	314	Y011_MYCGE	Hypothetical protein	0.01	287	1gsa_	Glutathione synthetase	–	> 10	
328	679	Y328_MYCGE	Hypothetical protein	0.01	756	1bgw_	Topoisomerase	–	> 10	
298	305	P115_MYCGE	P115 protein homolog	0.01	982	1sig_	RNA polymerase primary sigma	–	> 10	
087	71	RS12_MYCGE	30S ribosomal protein	0.01	139	1ah9_	Initiation factor 1 (if1)	–	> 10	
218	305	HMW2_MYCGE	Cytadherence	4×10^{-3}	1805	1sig_	RNA polymerase primary sigma	–	> 10	
029	501	Y029_MYCGE	Hypothetical protein	2×10^{-3}	186	1gpmA	GMP synthetase (xmp aminase)	–	> 10	
268	196	Y268_MYCGE	Hypothetical protein	1×10^{-3}	228	1ukz_	Uridylate kinase	1air	5.9	N
336	431	NISH_MYCGE	NIFS-like protein	6×10^{-4}	408	2dkb_	Decarboxylase	–	> 10	
057	378	Y057_MYCGE	Hypothetical protein	4×10^{-4}	178	1kay_	70 kDa heat shock protein	–	> 10	
062	514	PTFA_MYCGE	PTS system	3×10^{-4}	680	1occA	Cytochrome *c* oxidase	1occ	0.4	Y
353	96	Y353_MYCGE	Hypothetical protein	3×10^{-4}	109	1ihfA	Integration host factor	–	> 10	
133	514	Y133_MYCGE	Hypothetical protein	1×10^{-4}	228	1occC	Cytochrome *c* oxidase	–	> 10	
194	482	SYFA_MYCGE	Phenylalanyl-tRNA synthetase	1×10^{-4}	341	1lylA	Lysyl-tRNA synthetase (lysu)	1adj	5.8	Y
084	501	Y084_MYCGE	Hypothetical protein	6×10^{-5}	290	1gpmA	GMP synthetase (xmp aminase)	–	> 10	
050	311	DEOC_MYCGE	Deoxyribose-phosphate	4×10^{-5}	223	1dorA	Dihydroorotate dehydrogenase	–	> 10	
463	386	KSGA_MYCGE	Dimethyladenosine transferase	2×10^{-5}	259	2admA	Adenine-DNA-methyltransferase	1gcb	6.6	N
430	449	PMGI_MYCGE	2,3-Bisphosphoglycerate	1×10^{-5}	507	1alkA	Alkaline phosphatase	1alk	0.7	Y
372	501	Y372_MYCGE	Hypothetical protein	3×10^{-6}	385	1gpmA	GMP synthetase (xmp aminase)	–	> 10	
253	468	SYC_MYCGE	Cysteinyl-tRNA synthetase	2×10^{-6}	428	1gln_	Glutamyl-trna synthetase	–	> 10	
342	273	Y342_MYCGE	Hypothetical protein	2×10^{-6}	168	1qrdA	Quinone-reductase	–	> 10	
030	164	UPP_MYCGE	Uracil phosphoribosyltransferase	1×10^{-6}	206	1hgxA	Hypoxanthine phosphoribotransferase	1hgx	0.3	Y
347	293	Y347_MYCGE	Hypothetical protein	9×10^{-7}	210	1xvaA	Glycine n-methyltransferase	–	> 10	
423	228	Y423_MYCGE	Hypothetical protein	5×10^{-7}	561	1znbA	Metallo-β-lactamase	–	> 10	
094	303	DNAB_MYCGE	Replicative DNA helicase	8×10^{-9}	446	2reb_	Rec a protein	2reb	2.2	Y
126	317	SYW_MYCGE	Tryptophanyl-tRNA synthetase	2×10^{-9}	347	2ts1_	Tyrosyl-tRNA synthetase	–	> 10	
039	340	Y039_MYCGE	Hypothetical protein	1×10^{-9}	384	1an9A	D-amino acid oxidase	1an9	0.5	Y
394	426	GLYA_MYCGE	Serine hydroxymethyltransferase	4×10^{-10}	406	1tplA	Tyrosine phenol lyase	1gdo	0.3	N
186	135	Y186_MYCGE	Hypothetical lipoprotein	1×10^{-10}	250	1snc_	Staphylococcal nuclease	–	> 10	
333	273	Y333_MYCGE	Hypothetical protein	3×10^{-11}	126	1qrdA	Quinone-reductase	–	> 10	
412	321	Y412_MYCGE	Hypothetical lipoprotein	0	377	2abh_	Phosphate-binding protein	2abh	0.2	Y
273	678	ODPB_MYCGE	Pyruvate dehydrogenase E	0	326	1trkA	Transketolase	2trk	0.1	Y

The list of predictions obtained with the BASIC algorithm (see text) over and above predictions obtained with PSI-BLAST algorithm. BASIC, predictions obtained with the BASIC algorithm (see text); PSI-BLAST, predictions obtained with the PSI-BLAST algorithm [20]; MG, sequence number in the *M. genitalium* genome, as described in [23]; length, length of the protein (target or template); target name, name of the target protein, adapted from the SWISS-PROT 'DE' field; E value, E value (number of proteins with the same score expected by chance) of the best scoring template; PDB, Brookhaven PDB code; template name, name of the template protein, as present in the PDB file 'HEADER' field; A, agreement between BASIC and PSI-BLAST predictions: Y, both predictions agree; N, predictions disagree.

was postulated several years ago [29]. Recently, it was shown that the NIFS homologs are involved in the decomposition of several amino acids: a function similar to that of tyrosine lyase. All proteins from this superfamily show some sequence conservation around the PLP-binding site, but a large variety of active-site residues.

Third, serine methyltransferase (SWISS-PROT code GLYA_MYCGE, MG394) is predicted to belong to the same family as the NIFS protein, discussed above. This reaction is known to require pyridoxal phosphate, but this is a new reaction in this superfamily.

Fourth, deoxyribose phosphate aldolase (SWISS-PROT code DEOC_MYCGE, MG050) is predicted to have a TIM fold similar to that of other class I aldolases.

Fifth, dimethyladenosine transferase (SWISS-PROT code KSGA_MYCGE, MG459) is predicted to have a three-layered αβα fold similar to other DNA methylases.

Sixth, 2,3-biphosphoglycerate-independent phosphatase (SWISS-PROT code PMGI_MYCGE, MG426) is predicted to be similar to alkaline phosphatases.

Seventh, replicative DNA helicase (SWISS-PROT code DNAB_MYCGE, MG094) is predicted to have a structure similar to that of recA protein. Both proteins interact with single-stranded and double-stranded DNA with the helicase unwinding DNA during replication and recA catalyzing the pairing of homologous DNA sequences.

Finally, excinuclease ABC subunit C (SWISS-PROT code UVRC_MYCGE, MG204) is predicted to be structurally similar to 5′-exonuclease. This example will be analyzed in detail later.

In several of these predictions, it is possible to make an indirect confirmation of a PSI-BLAST result using a 'bridging protein' [30]. Such a protein is recognized as homologous to a prediction target and, at the same time, an independent search identifies its homology to another protein. Such predictions, however, have a rather low significance and require multiple BLAST runs. Here, they are identified in a single-step procedure and the prediction significance is high.

As mentioned earlier, predictions, such as those presented in Table 1, are difficult to verify unless a structure of a protein being predicted is determined experimentally. A comparison of the functions of both proteins, if known, can be used as an additional argument for or against their homology and, thus, indirectly verify their predicted structural similarity. Another possibility is to follow the structural prediction to its logical conclusion and build a three-dimensional model using the tools of competetive modeling. Building a three-dimensional model doesn't have strong predictive powers because misleadingly good models with the wrong topologies can be built, and sometimes otherwise correct models could not be built because of alignment errors [22]. Nevertheless, to illustrate such an approach, we have built the three-dimensional model of excinuclease ABC subunit C (SWISS-PROT code UVRC_MYCGE, MG204) using a T5 5′-exonuclease (PDB code 1exn; [31]) as a template. The model was built using the automated modeling program MODELLER [32] and the alignment was obtained from the BASIC program. The T5 5′-exonuclease has the unusual feature of a helical arch, which allows a single strand of DNA to thread through it [31]. Despite the very low sequence similarity of both proteins (13% of identical residues), the BASIC algorithm recognizes their similarity with a high E value of 0.03. An excinuclease model is presented in Figure 1. It is interesting to note that the unusual helical arch aligns very well with two predicted helices in 5′-exonuclease and a series of positively charged residues on the inside of the arch is perfectly reproduced in the model.

Functional predictions

The strong functional similarity between *M. genitalium* proteins with known functions, and their predicted structural 'cousins' is a strong argument that, as expected, the BASIC algorithm recognizes distantly related homologous proteins. Following this interpretation of the prediction results, there are several predicted relationships that provide new insights into the metabolism and other processes in *M. genitalium*. Three examples are now given.

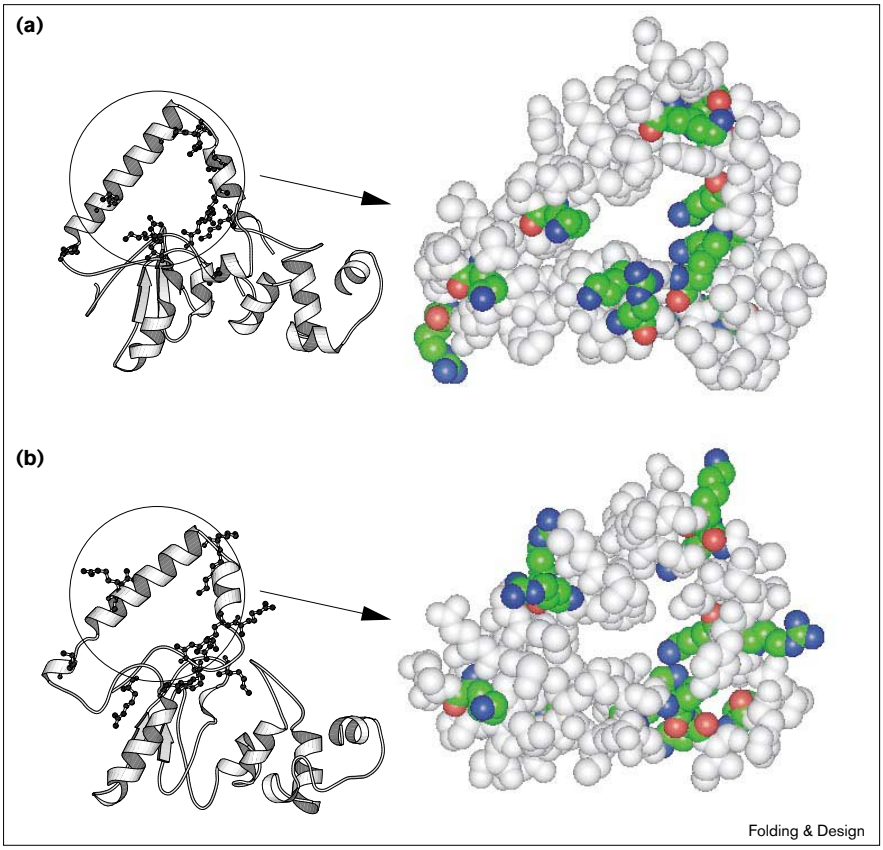
First, a second enzyme involved in amino acid metabolism is identified by the homology between hypothetical protein MG347, Y347_MYCGE (see Table 1) and glycine methyltransferase (PDB code 1xve).

Second, an additional enzyme involved in the synthesis of nucleic acid components is identified by the homology of hypothetical proteins MG372 (Table 2) and MG084 (Table 3) with GMP synthetase (PDB code 1gpm).

Finally, an intriguing homology is found for the hypothetical proteins MG423 (Table 2) and MG139 (Table 3) with β-lactamase. At first glance, this does not make sense because *M. genitalium* is not sensitive to penicillin or other β-lactam containing antibiotics (*M. genitalium* lacks a cell wall) and no penicillin-binding proteins were found in *Mycoplasma* [33]. On the other hand, the problem of antibiotic resistance in multi-organism infections is not very well understood and it could be speculated that *M. genitalium* is an opportunistic pathogen that degrades antibiotics as a part of symbiotic relationship with other pathogens.

It is interesting to note that from 28 new fold assignments reported recently for this genome [21], 18 were confirmed

Figure 1



A comparison of the model of the *M. genitalium* excinulease ABC subunit C (SWISS-PROT code UVRC_MYCGE) to that of the template recognized by the BASIC algorithm, the 5′-exonuclease from phage T5 (PDB code 1exn; [31]). **(a)** Structure of the template, T5 5′-exonuclease. **(b)** The excinuclease model prepared with the MODELLER [32] algorithm from the alignment obtained with the BASIC program. For both proteins, the structure of an unusual helical arch (circled in the ribbon diagram and enlarged as the space-filling model) and positively charged residues (ball-and-stick representation in the ribbon diagram and dark shading in the space-filling model) are shown. It is interesting to note that none of the individual residues is conserved, but the overall charge of this crucial structural fragment is identical in both proteins.

by a new generation of the BLAST algorithm, and an additional six were confirmed with high reliability and two with lower reliability by the method presented here. In only two cases from the list of 28 predictions of Fischer and Eisenberg [21], the BASIC algorithm did not recognize any above average similarity to a protein with a known structure. Several predictions from the list of below threshold predictions presented by Fischer and Eisenberg were also confirmed either by the new BLAST or by the BASIC algorithm.

In all cases discussed so far, homology to proteins with known structures was sought to make a fold prediction possible. There are, however, many proteins with functions determined by experiment, but without known structures. Establishing homology to such a protein family does not allow a structural prediction to be made, but allows some general predictions about the function of the new protein. Threading methods, that use structural information about one of the proteins to enhance the recognition and alignment accuracy can not be used in such application. To study how many distant relations could be recognized with more sensitive sequence similarity tools, a database of protein profiles was prepared for all proteins from the *Escherichia coli* genome. Structure and function predictions for proteins from this genome are described in a separate publication (L.R., B.Z. and A.G., unpublished observations). Here, only standard database annotations were used.

Of the 468 proteins from the *M. genitalium* genome, 96 are not homologous to any known proteins and were not, therefore, being annotated by the original authors. 56 proteins were similar to other proteins with unknown function and were, therefore, described as hypothetical proteins; 317 proteins had assigned function based on homology to an already characterized protein family. The PSI-BLAST calculations analyzed previously for recognition of proteins with known structures were now analyzed for recognition of *E. coli* proteins. The BASIC algorithm was used with a database containing sequence profiles of all *E. coli* proteins. The E value significance threshold of 0.05 for BASIC predictions was used as before. For PSI-BLAST predictions, the threshold was lowered to include any prediction with an E value < 10 (i.e. any prediction included in the standard PSI-BLAST output).

In the group of proteins with no annotations, 22 could be matched to other proteins from *E. coli*, 11 using PSI-BLAST and 22 using the BASIC algorithm. The results are presented in Table 2. As before, BASIC recognition is completely inclusive of the BLAST recognition, with the BASIC algorithm

Table 2

Homology assignments for orphan open reading frames from *M. genitalium*.

E value	Length M	Myco	Length E	Name of *E. coli* sequence		BLAST result
0.00457	251	MG116	236	GLMU_ECOLI	UDP-N-Acetylglucosamine pyrophosphorylase	
0.00237	109	MG353	99	IHFA_ECOLI	Integration host factor alpha-subunit	
0.00223	287	MG011	300	RIMK_ECOLI	Ribosomal protein S6 modification protein	B
0.00215	178	MG319	372	RFC_ECOLI	O-antigen polymerase	
0.00077	420	MG181	427	ARSB_ECOLI	Arsenical pump membrane protein	
0.00063	393	MG306	509	NUOM_ECOLI	NADH dehydrogenase I chain M	
6.6×10^{-5}	118	MG436	148	YCCF_ECOLI	Hypothetical 16.3 kDa protein	
4.2×10^{-6}	561	MG423	215	YCBL_ECOLI	Hypothetical 23.8 kDa protein	B
3.9×10^{-6}	140	MG236	191	YJBK_ECOLI	Hypothetical 21.7 kDa protein	
1.1×10^{-10}	141	MG427	143	OSMC_ECOLI	Osmotically inducible protein	
1.2×10^{-11}	196	MG208	231	YGJD_ECOLI	Hypothetical 36.0 kDa protein	
0	128	MG207	183	YFCE_ECOLI	Hypothetical 20.1 kDa protein	B
0	168	MG342	188	YIEF_ECOLI	Hypothetical 20.4 kDa protein	B
0	483	MG045	370	POTF_ECOLI	Putrescine-binding periplasmic protein precursor	B
0	126	MG333	201	ACPD_ECOLI	Acyl carrier protein phosphodiesterase	B
0	144	MG449	110	YGJH_ECOLI	Hypothetical 12.3 kDa protein	B
0	377	MG412	346	PSTS_ECOLI	Phosphate-binding periplasmic protein precursor	B
0	236	MG385	247	UGPQ_ECOLI	Glycerophosphoryl diester phosphodiesterase	B
0	385	MG372	482	YAJK_ECOLI	Hypothetical 55.0 kDa protein	B
0	291	MG468	280	EXO_ECOLI	Potential 5′–3′ exonuclease	B
0	557	MG369	210	ORF 01172		
0	153	MG230	136	NRDI_ECOLI	NRDI protein	B

E value; number of hits that could be obtained with this score by chance; Length M, length of *Mycoplasma* sequence; Myco, name of *Mycoplasma* sequence; Length E, length of *E. coli* sequence; B, confirmed.

identifying 11 new proteins. Of the 22 proteins, seven are matched with hypothetical proteins; thus, no functional prediction is possible. For the remaining 15 proteins (six from the group identified by both algorithms and nine from the group identified entirely by BASIC), tentative functional assignments could be made based on their classification into an already characterized homologous family.

For 56 hypothetical proteins from the *M. genitalium* genome, 14 can be assigned to *E. coli* proteins with known

Table 3

Homology assignments for hypothetical proteins from *M. genitalium*.

E value	Length M	Myco	Length E	Name of *E. coli* sequence		BLAST result
0.03566	343	MG205	112	GATR_ECOLI	Galactitol utilization operon repressor	
0.01703	443	MG314	57	RL32_ECOLI	50S Ribosomal protein L32	
0.00477	178	MG057	581	PRIM_ECOLI	DNA primase	
0.00031	306	MG121	336	MGLC_ECOLI	Galactoside transport system permease protein	B
6.1×10^{-5}	280	MG135	46	RL34_ECOLI	50S Ribosomal protein L34	
4.4×10^{-5}	425	MG461	505	DGTP_ECOLI	Deoxyguanosinetriphosphate triphosphohydrol	B
3.2×10^{-7}	569	MG139	252	PHNP_ECOLI	PHNP protein	B
2.8×10^{-7}	227	MG323	458	TRKA_ECOLI	TRK system potassium uptake protein	B
3.6×10^{-10}	324	MG371	75	FEOA_ECOLI	Ferrous iron transport protein	
0	489	MG225	461	YIFK_ECOLI	Probable transport protein	B
0	290	MG084	432	MESJ_ECOLI	Cell cycle protein	B
0	385	MG464	548	60IM_ECOLI	60 kDa inner-membrane protein	
0	448	MG329	503	THDF_ECOLI	Thiophene and furan oxidation protein	B
0	323	MG370	326	SFHB_ECOLI	SFHB protein	B
0	112	MG143	133	RBFA_ECOLI	Ribosome-binding factor A (P15B protein)	B
0	239	MG247	217	FDNI_ECOLI	Formate dehydrogenase, nitrate-inducible	B
0	336	MG270	338	LPLA_ECOLI	Lipoate-protein ligase	B

E value, number of hits that could be obtained with this score by chance; Length M, length of *Mycoplasma* sequence; Myco, name of *Mycoplasma* sequence; Length E, length of *E. coli* sequence; B, confirmed.

function, with 12 of them assigned by PSI-BLAST. There are also three proteins whose function is known in *M. genitalium* that are homologous to hypothetical proteins from *E. coli*. All thus identified proteins are listed in Table 3. In the latter case, all pairs are recognized both by PSI-BLAST and the BASIC algorithm. The complete list of the comparison of the *M. genitalium* proteins to the *E. coli* genome is presented on the authors' World Wide Web pages.

The analysis presented above was designed to compare relative sensitivities of PSI-BLAST and BASIC algorithms. It is by no means a complete analysis of function assignments possible for uncharacterized *M. genitalium* proteins. Such assignments could be done with a more complete analysis of PSI-BLAST output and/or with a further increase of the database of sequence profiles used by the BASIC algorithm. Such an analysis is currently in progress and will be the subject of a separate publication.

Discussion

The identification of distant evolutionary relationships is currently the most reliable structure and function prediction tool. The position-specific iterative BLAST algorithm represents the most sensitive of the widely available algorithms for such identification. For instance, it was shown here that this algorithm can assign folds to 25% of *M. genitalium* proteins, including most of the new predictions obtained using the 3D1D threading algorithm of Fischer and Eisenberg [21]. The PSI-BLAST algorithm achieved its high level of prediction accuracy by accounting for different mutation rules at different positions by automatically creating a sequence profile from a set of close homologs.

A new BASIC algorithm takes it one step further and compares a profile to a database of precalculated protein profiles. It enabled us to identify 50 additional homologies between proteins from the *M. genitalium* and well-characterized protein families, bringing the total number of fold assignments to 176, or 38% of the entire genome. This represents an increase of > 70% over recent threading-based fold assignments and an almost 50% increase over the latest generation of the BLAST algorithm. This is a conservative estimate because rather stringent significance criteria were used to identify the BASIC predictions.

One has to bear in mind, however, that the prediction significance, as calculated by PSI-BLAST or BASIC algorithms is based on comparing the alignment score to the distribution of scores for the entire database. It is possible that the score differs from all the other scores for a reason other than the homology of the two proteins. For instance, an unusual composition of the prediction target may result in a 'significant' score to another protein with similar amino acid composition, despite a lack of any relationship between the two proteins. It is possible that in this sense high-significance prediction might be incorrect, even though we have failed to find such a case so far. Several strong predictions of similarity to RNA polymerase (PDB code 1sig) or topoisomerase (PDB code 1bgw) are possible exceptions. In these two proteins there are long fragments of coiled-coil structure, which can be matched to coiled-coil regions from other proteins, possibly without any homology between them. In this case, PSI-BLAST often predicts strong similarity to tropomyosin. Thus, prediction results, such as presented in Table 1, must always be interpreted with caution and other factors, such as the similarity between the functions of both proteins (if known), should be taken into account when evaluating possible homology between assignments in Table 1.

All predictions presented here and on the authors' World Wide Web site, represent genuine structural and functional predictions and, as such, are difficult to verify. By making them public, we invite verification of our predictions by experiment and other prediction algorithms. In all cases for which the function of a protein whose structure was predicted was known, however, it was possible to identify some analogy between this function and the functions of proteins from the homologous family identified in the prediction.

Despite the experimental status of the BASIC algorithm and other fold and function prediction algorithms, the actual predictions presented here illustrate the practical importance of such analyses. They provide circumstantial evidence that, if used in conjunction with other data, can offer deep insights into the function of partly characterized gene products. The most interesting situation arises when the general function is known from, for instance, knockout genomic experiments. Structural and functional predictions such as those presented here provide hints about mechanisms and activities of specific proteins that are involved in this function.

Because both PSI-BLAST and BASIC algorithms do not use information about protein structure, both can be applied to search for homologs among proteins with known functions, but without known structures. To compare both algorithms in this task, the proteins from *M. genitalium* genome were compared to those from the *E. coli* genome. When compared to annotations available for the *M. genitalium* genome from the World Wide Web site for The Institute for Genome Research, 40 additional homologies were identified, with 16 of them recognized only by the BASIC algorithm. 26 proteins without known homologs were assigned to *E. coli* families and for 16 of them, a tentative function prediction could be made. In addition, for 14 hypothetical proteins with their only known homologs from the uncharacterized open reading frame from other genomes, homologies to already characterized protein families were found. Again, these are conservative estimates, because very strict significance thresholds were used.

Materials and methods

PSI-BLAST and the sequence database

The position-specific iterative BLAST algorithm [20] is the newest version of the *de facto* standard of database protein similarity searching algorithms. This algorithm addresses the principal shortcoming of the previous BLAST algorithm: its inability to introduce gaps in the alignment. In addition, the PSI-BLAST algorithm allows the iterative building of a sequence profile from the multiple alignment of sequences of homologous protein identified in the first pass of the algorithm. The PSI-BLAST program was downloaded from the NIH World Wide Web site (National Center for Biotechnology Information, URL: www.ncbi.nlm.nih.gov) and used following the guidelines in the manual. The sequence database used by the PSI-BLAST algorithm contains a non-redundant compilation of sequences available from SWISSPROT and PIR databases, as well as translated DNA sequences from the EMBL and NCBI nucleotide sequence databases and sequences of all proteins deposited in the Brookhaven PDB. This database was used to prepare sequence profiles (see below) for all targets and all templates and, thus, is a complete super-set of the database used by the BASIC method. The version used in this work was compiled in November 1997.

Profile sequence preparation

The method described in this paper is based on an evaluation of the similarity between two sequence profiles. A sequence profile is a position-specific probability distribution, which for every position along the sequence gives a probability that one of the 20 amino acids would occupy this position [12,13]. Profiles were generated automatically using the multiple alignment of homologous sequences as generated by the PSI-BLAST algorithm. The technical details of the profile preparation are described in a separate publication [14]. Exactly the same procedure is followed for the target proteins as for all proteins contained in the databases being searched.

Databases of sequence profiles

Two databases were constructed for the work described here. The first database of 1151 representative protein structures was prepared on the basis of a non-redundant set of protein structures included in the FSSP database as available from the DALI server at EBI. This database was used for fold prediction. The second database consists of sequence profiles for all proteins from the *E. coli* genome, as available on the *E. coli* World Wide Web site at the University of Wisconsin Genome Center (URL: www.genetics.wisc.edu).

The BASIC profile-to-profile alignment algorithm

Two sequence profiles are compared in the same way as two sequences. A local–local version of a Smith-Waterman dynamic programming algorithm is used [34]. The similarity score between positions in two sequences is, however, calculated with the mutation matrix, such as for the Gonnett similarity matrix [35]. For two profiles, this value is calculated as an average of scores between all amino acid pairs, averaged according to the probability distribution in each profile. Three parameters, gap introduction penalty, gap extension penalty and a constant, added to each element of the mutation matrix, are optimized for a fold recognition benchmark, as described below.

Optimization and verification of the BASIC algorithm

The BASIC algorithm was optimized to recognize the maximal number of structurally similar proteins on benchmarks customized for fold-prediction algorithms. A particular benchmark available from the World Wide Web server at UCLA (URL: fold.doe-mbi.ucla.edu) was used during the development of a BASIC algorithm. This benchmark consists of 68 target proteins for which the correct template (structurally similar protein) has to be found in a database of ~300 examples. The results (Table 4) presented here show that a sequence-only fold recognition method can closely match the prediction accuracy of best threading algorithms. A more extensive evaluation of different fold recognition algorithms is presented elsewhere.

Table 4

Fold prediction accuracy of different algorithms.

Program	Rank = 1	Rank ≤ 5	Rank ≤ 10
Simple BLAST	27	–	–
PSI-BLAST	32	–	–
BASIC THREADING	22	30	34
Global sequence alignment	40	50	52
Hybrid THREADING	54	58	60
BASIC	52	57	60

Results achieved on the UCLA threading benchmark with 68 target–template pairs and a database of 300 templates. The values given are the number of pairs for which the template obtained the rank indicated. For BLAST predictions, it is difficult to estimate lower significance predictions because often they are not listed because of a large number of homologous proteins. BLAST, BLAST 1.02 version [1]; PSI-BLAST, [20]; BASIC THREADING, threading as described in [9]; global sequence alignment, sequence alignment using the GONET substitution matrix and the global–local dynamic programming subroutine; hybrid THREADING, hybrid threading version [22]; BASIC, method presented here (local alignment).

Score significance

Scores of individual profile–profile comparisons are corrected for the size of the proteins being compared [34,36] and used to calculate the distributions of scores for a given prediction target. The empirical distribution was fitted to an extreme value distribution. The parameters of this fit were used to calculate the E value, i.e. the expected number of proteins with the given score in a given database.

The estimation of the reliability of the prediction was based on the E value statistic. The cutoff of 0.05 for the E value used here is much bigger then the scores of false positive answers of the procedure observed during the development. The biggest E value for a false positive in the UCLA benchmark described above was 1.6. At this point, however, it is not known how much the distribution of scores on the training set is different from the distribution on the larger set used in the actual predictions. For this reason, we use a very conservative significance threshold.

A version of the BASIC program is available on the group's World Wide Web site. It offers the possibility of similarity predictions in the database of structural families, as described above. The user can supply the sequence of the target protein.

Acknowledgements

The authors are grateful for many stimulating discussions with A. Kolinski, J. Skolnick and J. Fetrow. We also thank A. Šali for providing us with a MODELLER software package, as well as for help and advice in using it. This work was supported by NIH Grant No. GM48835.

References

1. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403-410.
2. Pearson, W.R. & Miller, W. (1992). Dynamic programming algorithms for biological sequence comparison. *Methods Enzymol.* **210**, 575-601.
3. Wisconsin Package Version 9, Genetics Computer Group (GCG), Madison, Wisconsin, USA. World Wide Web URL: www.gcg.com.
4. Geourjon, C. & Deleage, G. (1995). ANTHEPROT 2.0: a three-dimensional module fully coupled with protein sequence analysis methods. *J. Mol. Graphics* **13**, 209-212.
5. (1996). SWISSMODEL, an automated knowledge based modelling server. World Wide Web URL: http://expasy.hcuge.ch/swissmod/swiss-model.html.
6. Henikoff, S., Pietrokovski, S. & Henikoff, J.G. (1998). Superior performance in protein homology detection with the BLOCKS database server. *Nucleic Acids Res.* **26**, 309-312.
7. Bowie, J.U., Luethy, R. & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three dimensional structure. *Science* **253**, 164-170.

8. Jones, D.T., Taylor, W.R. & Thornton, J.M. (1992). A new approach to protein fold recognition. *Nature* **358**, 86-89.
9. Godzik, A., Skolnick, J. & Kolinski, A. (1992). A topology fingerprint approach to the inverse folding problem. *J. Mol. Biol.* **227**, 227-238.
10. Sippl, M.J. & Weitckus, S. (1992). Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a database of known protein conformations. *Proteins* **13**, 258-271.
11. Bryant, S.H. & Lawrence, C.E. (1993). An empirical energy function for threading protein sequence through folding motif. *Proteins* **16**, 92-112.
12. Gribskov, M., McLachlan, M. & Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. *Proc. Natl Acad. Sci. USA* **84**, 4355-4358.
13. Bork, P. & Gibson, T.J. (1996). Applying motif and profile searches. *Methods Enzymol.* **266**, 162-184.
14. Rychlewski, L., Jaroszewski, L., Zhang, B. & Godzik, A. (1998). Comparison of sequence profiles. Structural predictions with no structure information. *Protein Sci.*, in press.
15. Pastore, A. & Lesk, A.M. (1990). Comparison of the structure of globins and phycocyanins: evidence for evolutionary relationship. *Proteins* **8**, 133-155.
16. Holm, L. & Sander, C. (1995). Evolutionary link between glycogen phosphorylase and a DNA modifying enzyme. *EMBO J.* **14**, 1287-1293.
17. Babbitt, P.C., *et al*., & Gerlt, J.A. (1995). Functionally diverse enzyme superfamily that abstracts the α proton of carboxylic acids. *Science* **267**, 1159-1161.
18. Murzin, A.G. & Bateman, A. (1997). Distant homology recognition using structural classification of proteins. *Proteins* **Suppl. 1**, 105-112.
19. Karplus, K., *et al*., & Sander, C. (1997). Predicting protein structure using Hidden Markov Models. *Proteins* **Suppl. 1**, 134-139.
20. Altschul, S.F., *et al*., & Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acid Res.* **25**, 3389-3402.
21. Fischer, D. & Eisenberg, D. (1997). Assigning folds to the proteins encoded by the genome of *Mycoplasma genitalium*. *Proc. Natl Acad. Sci. USA* **94**, 11929-11934.
22. Jaroszewski, L., Rychlewski, L., Zhang, B. & Godzik, A. (1998). Fold prediction by a hierarchy of sequence and threading methods. *Protein Sci*., in press.
23. Fraser, C.M., *et al*., & Ventor, J.C. (1995). The minimal gene complement of *Mycoplasma genitalium*. *Science* **270**, 397-403.
24. Moult, J. (1996). The current state of the art in protein structure prediction. *Curr. Opin. Biotechnol.* **7**, 422-427.
25. Francklyn, C., Musier-Forsyth, K. & Martinis, S.A. (1997). Aminoacyl-tRNA synthetases in biology and disease: new evidence for structural and functional diversity in an ancient family of enzymes. *RNA* **3**, 954-960.
26. Lodmell, J.S. & Dahlber, A.E. (1997). A conformational switch in *Escherichia coli* 16S ribosomal RNA during decoding of messenger RNA. *Science* **277**, 1262-1267.
27. Sette, M., *et al*., & Boelens, R. (1997). The structure of the translational initiation factor IF1 from *E. coli* contains an oligomer binding motif. *EMBO J.* **16**, 1436-1443.
28. Hubbard, T.J.P., Murzin, A.G., Brenner, S.E. & Chothia, C. (1997). SCOP: a structural classification of the protein database. *Nucleic Acids Res.* **25**, 236-239.
29. Ouzounis, C., Sander, C., Scharf, M. & Schneider, R. (1993). Prediction of protein structure by evaluation of sequence-structure fitness: aligning sequences to contact profiles derived from 3D structures. *J. Mol. Biol.* **232**, 805-825.
30. Pearson, W.R. (1997). Identifying distantly related protein sequences. *Comput. Appl. Biosci.* **13**, 325-332.
31. Ceska, T.A., Sayers, J.R., Stier, G. & Suck, D. (1996). A helical arch allowing single-stranded DNA to thread through T5 5'-exonuclease. *Nature* **382**, 90-93.
32. Šali, A., Potterton, L., Yuan, F., van Vlijmen, H. & Karplus, M. (1995). Evaluation of comparative protein modeling by MODELLER. *Proteins* **23**, 318-326.
33. Martin, H.H. (1980). Differentiation of mycoplasmalates from bacterial protoplast L-forms by assay for penicillin binding. *Arch. Microbiol.* **127**, 297-299.
34. Waterman, M.S. (1995). *Introduction to Computational Biology: Maps, Sequences and Genomes (Interdisciplinary Statistics)*. Chapman & Hall, New York.
35. Gonnet, G.H., Cohen, M.A. & Benner, S.A. (1992). Analysis of amino acid substitution during divergent evolution. *Science* **256**, 1443-1445.
36. Karlin, S. & Altschul, S.F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl Acad. Sci. USA* **87**, 2264-2268.

Because *Folding & Design* operates a 'Continuous Publication System' for Research Papers, this paper has been published on the internet before being printed. The paper can be accessed from http://biomednet.com/cbiology/fad – for further information, see the explanation on the contents pages.